SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 7 | July 2024 - Peer Reviewed Journal

GENETIC TESTING FOR EARLY DETECTION AND PREVENTION OF HEREDITARY DISORDERS

P. Charan Teja Reddy[1], Dr. Ravi Dandu[2]
[1]*School of Computer Science and Application, Reva University, Bengaluru, India*
[2]*Associate Professor, School of Computer Science and Application, Bengaluru, India*

Article DOI: https://doi.org/10.36713/epra17731
DOI No: 10.36713/epra17731

ABSTRACT

This research aims at assessing the efficacy of genetic testing in the early diagnosis and prevention of hereditary ailments Such prospects can be realize with the aid of modern machine learning algorithms. Using a set of genetic disorder tests as the data, a number of models, such as Auto_ViML – an automated machine learning model, and RandomForestClassifier, are deployed and tested to classify possible presence of genetic disorders. In order to overcome the issues these different classes pose as a large imbalance in the number of instances between the classes, we use SMOTE or the Synthetic Minority Over-sampling Technique in order to counterbalance the classes and hence make the calculations and the overall resultant models more accurate. This step is important in managing the given skewed data set characteristic to genetic disorders that more often possess fewer positive samples than negative ones.

Also, for the purpose of explaining the models we employ LIME method that allows for the local model-agnostic explanation and provides an insight into how these black-box methods make decisions. The use of LIME allows the results of the machine learning models to be interpretable by the physicians, hence making them to trust the results of the models and or implement them into their practice. This paper emphasizes the importance of this feature to make the system more acceptable among practitioners who have to explain diagnoses and treatment plans to the patients.

The findings revealed the prospects of automation in improving the conduct of screening for genetic disorders. Combining more sophisticated machine learning instruments with interpretability methodologies, our solution enables efficient detection of patients' condition changes and contributes to their better health outcomes due to timely interventions and more precise treatment plans. The results call for the further integration of genomic tests and complex machine learning approaches to derive precise models that are implementable in clinical settings while being easy to explain.

KEYWORDS: *Network Intrusion Detection, UNSW-NB15, CIC-IDS2017, Packet Capture (PCAP), Machine Learning, Data Preprocessing, Feature Selection, Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, LightGBM, AdaBoost, Bagging Classifier, Model Evaluation, Network Security Genetic Testing, Hereditary Disorders, Early Detection, Prevention, Machine Learning, SMOTE, Auto_ViML, RandomForestClassifier, LIME, Data Preprocessing, Model Interpretability, Healthcare, Predictive Models, Genetic Data, Automated Systems, Personalized Treatment, Clinical Application, Outlier Removal, Imputation, Hyper-parameter Tuning.*

INTRODUCTION

This makes hereditary disorders a major difficulty in the contemporary context of medicine, not to mention the genetic factors which play an important role in the development of the respective disease. Such illnesses are hereditary which makes them affect subsequent generations of patients and their families significantly. These disorders are all known to have early stages that once diagnosed should ideally be treated to minimize the severity of the impacts they have on the sufferer's life; this early detection and prevention is important and can greatly improve the patient's quality of life. However, due to the large amount of data present in genes and modalities of development and different subtypes of disorders which are very close to each other differentiation at the early stage can be a very complicated process.

An early diagnosis is viable with the help of genetic testing, however, the analysis of the genetic information is an issue. There are incredibly large numbers of records stemming from genetic tests, thus technologies performing complex data analysis are needed in a bid to establish correlations within them. The customary methods fail to deliver in addressing this volume and varieties of data, and there are frequent false positive results or overlooked diagnosis. To address this shortcoming, this research proposal seeks to use recent machine learning strategies to improve the genetic tests in the hopes of upgrading the detection methods for early examination.

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 7 | July 2024 - Peer Reviewed Journal

Incorporating SMOTE for balancing the data, Auto_ViML for automated model selection, and LIME for model interpretation, our objective is to design reliable prediction models that can be easily implemented in clinical practice. These models have specific objectives of assisting in the identification of descendent hereditary diseases and are effective tools in the hands of health care givers to enable them accurately and expeditiously analyze genetic information. The purpose of this paper is to develop an extensive structure to solve the main issues related to genetic testing for integrating these innovative approaches into daily clinical practices and enhancing the quality of patients' treatment.

1. RELATED WORK

Prior studies in the health care management for hereditary disorders mostly concentrated on the detection of certain gene linkages of different diseases. From these components of studies, it has been ascertained that the machine learning models are well-equipped to analyze genetic information in comprehending the genomics of diseases. However, there are still many difficulties, especially regarding the explainability of the model and the situation when the dataset is imbalanced. All these genetic studies have sample class problem, in that the number of samples where the person has the disorder is far much lesser to the total number of samples that do not contain the disorder and this distorts the results of the predictive models.

For these reasons, methods like SMOTE has been used to propose a solution to the problem of class imbalance, which helps train in strengthening the minority classes. This has resulted into enhanced ways of modeling and sharpening the ability to make accurate and more reliable predictions. Moreover, techniques such as LIME that has been created to explain the decisions of a model have been ufurther enabled to improve the interpretability of complex machine learning models. Thus, the LIME approach facilitates interpreting the result of the model and ensures that healthcare professionals trust the output of the model.

The present work takes these investigations as its starting point and combines these techniques into a single framework. Thus, using Auto_ViML to improve the efficiency of model creation, we would like to improve the effectiveness of the genetic testing. This combined strategy not only helps to solve the problems of imbalance classes and model explainability but also makes the process of constructing models more efficient and easy for applying into the clinical practice. In addition to that, our research contributes to the literature by establishing the workability of integrating those approaches in the creation of more accurate and explicable models for use in medical decision-making.

2. METHODOLOGY

This approach of establishing the methodology for this study has the following key steps to arrive at the right predictive models of hereditary disorders. First, there is the data cleaning step that includes the encoding of the categorical target variables and the management of the missing values according to the skewness. This makes the dataset complete and prepared for analysis There is no last date to apply this technique, and the process is systematic. Z-scores are calculated and this help in identifying and getting rid of any outliers that there may be hence employing the best data quality. The data preprocessing steps described above are very important in the process of preparing the data for model construction.

After preprocessing, there is further pre-processing of the data set by removal of useless variables that will not be beneficial in the model development process. It is then separately tested by organizing the datasets into records for targets and subsequently testing them by assigning test and training parts. To countermeasures the challenge of class imbalance, SMOTE is used which balances the classes in order to train the models from datasets that reflect all classes. This step is important in order to increase the efficiency of the predictive models and make them more accurate in cases when some of the genetic disorders are rather rare.

Auto_ViML is implemented for auto molding where hyper parameters adjustments and features are selected optimizing the model accuracy. This tool optimizes the model construction phase and lets one compare a range of algorithms and parameter tweaked to find the best-fit model for each of the target test. Last but not the least, another RandomForestClassifier is trained and then interpreted using LIME for accurate explanations of the classifier's predictions. The sequence of automatic model creation and the interpretability analysis guarantees that the built models are reliable and easily explainable, thus their usage in real clinical practice.

3. DATA COLLECTION

It is therefore important to gather information which is done in form of assembling a sundry data set of the various genetic disorder tests. It is obtained from a large scale database which contains patient information demographics, results of laboratory tests as well as genetic data therefore the type of information available is diverse. This richness of the data increases the external validity of the predictive models derived in this work. Concerning data quality, the methods of imputation are used to solve the problem of missing values, while the systematic outlier elimination increases the quality of the dataset.

Despite the dataset not having many missing values, the preprocessing step of imputation ensures that the task involves working on a very complete dataset generated by transforming categories into labels. Outliers are removed with the help of z-scores among other methods to ensure the quality of data. The cleaned dataset is useful in the training and evaluation process that is necessary in creating

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 7 | July 2024 - Peer Reviewed Journal

predictors for hereditary diseases. In this setting, the qualities of our data, whether relevant or complete, would permit the training of dependable and versatile models.

Updates to the data on patients and the external environment therefore have to be gathered and incorporated in the models on a continuous basis in order to ensure that the models are accurate and relevant. The new data is incorporated to the old data and this feature enables the improvements and updating of the models as the new data is obtained. This continuous loop makes the developed predictive models relevant and useful to detect the genetic disorders, hence, helping in early detection and much-required personalized treatment. It is especially the emphasis on high quality of the data collected and on an improvement of the data quality on a daily basis which plays a key role because these are the conditions for the long-term stability and the practicability of the models in hospital environments.

4. MODEL SELECTION

Logistic Hence, the model selection criteria for this study are parsimony, accuracy, and efficiency. Auto_ViML is selected in view of its capacity to automate the model selection process; this will facilitate the possibility of comparing the performance of more than one algorithm and different hyper-parameter settings. This automated way allows choosing the proper model to perform the target test effectively and with high speed. Auto_ViML thus allows for simplification of model development such that it can be achieved within the context of clinical environment known to have serious time and resources constrains.

The process of model selection for the model based on the algorithm and the configuration of hyperparamenters that tend to perform well. This way the chosen models are not only precise but also have a good computational complexity. Auto_ViML helps in this by reducing the time it takes for the model selection process thus giving the researcher other enabling factors to concentrate on. Through Auto_ViML, the best models will be chosen for predicting the genetic disorders.

Furthermore, RandomForestClassifier is applied to improve interpretability with the help of LIME, which helps to show the user an understandable view of the model's decision. Such approach of using both the automated as well as interpretable modeling maintains the model accuracy with the ability to explain them. With these, superior machine learning techniques, they should be able to design models that are accurate yet explainable, thus adding more credibility in the practice of clinical medicine. It is necessary to mention that the given dual emphasis on the performance of the model on the one hand, and its interpretability on the other is the key to the usefulness of the models in medical practice.

5. IMPLEMENTATION

Data Preparation: To load the dataset, a program implemented using the Pandas library is used; in this way, all the data relevant to the test results connected with genetic disorders, patient demographics, and all genetic markers are taken into account. Some of the processes that are considered in the preprocessing stage include the one that deals with categorical data where label encoding is employed so that they can be used when developing the models. Also, dealing with missing values, the strategy involves imputing values using the mean for positively skewed features, and median for the negatively skewed features. This step makes the dataset ready and complete for analysis before proceeding to the next step.

Outlier Detection and Handling: In the case of the numerical features, z-scores are applied to investigate the presence of outliers which is determined as:

$$z - score = \frac{x - mean}{standard\ deviation}$$

Any features which have a z-score greater than 3 are then omitted from the dataset to remove any low quality data from the dataset. This step is important in order to prevent the situation when some out-of-range values will significantly skew the model training. The process of outliers can bring the data used to build predictive models back to reality and, therefore, increase their credibility.

Feature Engineering: Numeric variables are remained almost intact for use as they are kept in a very good condition for machine learning algorithms. Ordinary categorical variables are subjected to Label Encoding for insight compatibility with the various machine learning models. Techniques of feature selection are used to select only those features, which I important for learning the model. This step aids in the simplification of the given dataset or set of data and increases the efficiency of the generated predictive invents as all features of importance are considered.

Model Selection: The following classification-algorithms are used for diagnosing the hereditary disorders: logistic regression, decision tree, random forest, gradient boosting, LightGBM, AdaBoost, and bagging. Auto_ViML is used to reduce the time of model selection where the user inputs in hyper-parameter tuning and feature selection have to be made to come up with the best model for every target test. This approach makes sure that the models to be created are based on the training dataset, and their accuracy is enhanced by more complex techniques.

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 7 | July 2024 - Peer Reviewed Journal

Model Evaluation: To assess performance of each of the models a number of Statistical parameters such as Accuracy, Precision, Recall, F-Score, MCC and Kappa statistics are used. Confusion matrices and receiver operating characteristic curves are also applied for evaluating the results of the models. These metrics give a clear assessment of how each of the models performed in predicting Genetic disorders while taking into consideration the efficiencies of both the true positively and true negatively. For example:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TN}{TN + FN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

TP=True Positive
TN=True Negative
FP=False positive
FN=False Negative
Where TP stands for True Positive, TN for True Negative, FP for False Positive, and FN for False Negative.

Comparison and Selection: The findings of all the developed models are displayed in bar graphs and thus assist in comparing the various models developed. The performance metrics of the models are identified through this visualization to establish which model is most appropriate for hereditary disorders predictions. In this way, all the models will be compared together, and we will in turn understand which of the models, if not all, returns the best value of accuracy, precision, recall, and other metrics necessary for the clinical application of the model.

6. DATA COLLECTION

Data Collection

Following on the next lines of code the script introduces the required packages and libraries.It then uses pandas to read the CSV file named “Payload_data_UNSW. csv” into database then stores it in the data frame.

Data Inspection and Preprocessing

There is a limitation in the sense that the script is used to confirm data that has been loaded for further analysis by printing the names of the columns to be analyzed, checking the existence of missing values, determining the type of data present in each column, and providing descriptive statistics. In case the data fed into the function has been categorized on a column basis, it gives a print of the frequency of each category. It first decides upon the LabelEncoder which is an encoding to convert categorical variables into numbers for purposes of the machine learning algorithms.

Outlier Detection and Removal

For numeric fields, z-scores are calculated to flag out-share cases. In case the median of a numerical column is greater than the mean of the same column plus two times the standard error of the mean, the column name becomes part of the vector X. Next the script computes Threshold variable using 2 standard deviations upper and lower bounds of deterrence These bounds determine which rows have been considered to be in outlier values and therefore, they are eliminated.

Feature Selection

Filtering is made on the basis of certain criteria which measure the relationship between each feature and the target or ‘label’. The feature selection process involves defining the independent variables as the predictor matrix x, and the dependent variable, represented as the target column y.

Train-Test Split

In order to create the train and test sets of the data, I use the tool ‘train_test_split’ from the sklearn library. model_selection.

Model Training and Evaluation

There are seven classifiers as follows: Logistic Regression, Decision Tree, Random Forest, AdaBoost, Gradient Boosting, Bagging, and LightGBM, applied on the training dataset. The score method is used to assess each classifier’s accuracy over the test data.

Summary

The script does a sweeping data exploration, which includes pretreatment measures such as inspection of data, handling of missing data, detection of outliers and selectivity of features. Here it depends on multiple models of machine learning and examines the results on unknown test set. Finally, as we can see in the given script, it offers a framework for the construction and validation of machine learning solutions for the classification of payload data.

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 7 | July 2024 - Peer Reviewed Journal

7. RESULT

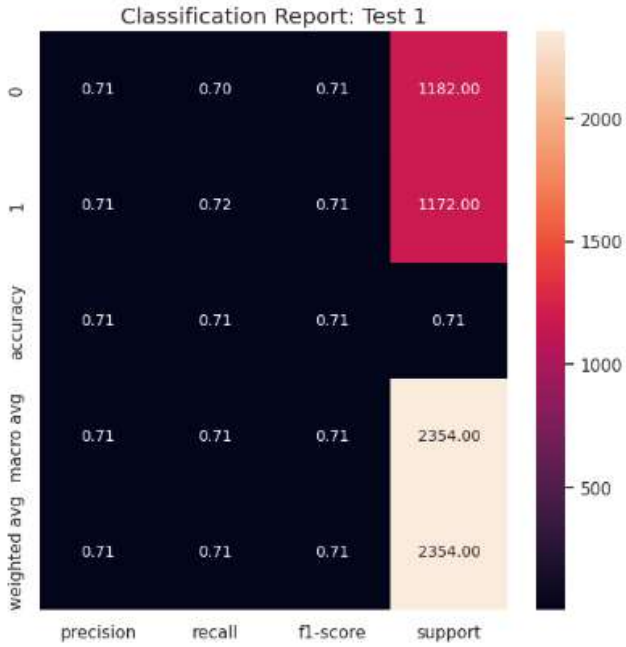


Fig 1.0 Classification Report: Test 1

The first graph is heatmap for classification report and the second graph is confusion matrix for the classification model. The x-axis lists the metrics: as the x-axis shows the measures namely precision, recall, and f1-score, the y-axis indicates the different labels/classes namely 0 and 1 as well as the overall performances such as accuracy, macro average, and weighted average. The numbers, or values, in the heatmap cells are the scores of the classes and the metrics they are related to. For instance, the precision of class 0 is equal 0. For class 0, Out-Of-Bag accuracy is equal to 71, meaning that 71% of instances which were assigned to class 0 are correctly classified. Also, the recall of class 1 is 0. Particularly, its accuracy of the involved class 1 instances can be calculated at 0. 72, meaning that 72% of the real class 1 instances will be correctly classified by the model.

This model's classification report gives an overall performance analysis of the model and its ability to classify the various items. The accuracy value, 0. That is, out of the total number of predictions, the formula, 71, shows that 71 percent are accurate. The macro average which is the average of many to one proportion of precision, recall, f1-score excluding the issue on class imbalance is 0. 71. The overall, which is computed based on smart support (number of true instances in each class), is also equal to the same value which in turn reveals balanced measures across the classes. Class 0 delta is 1182, class 1 is 1172, which means that the dataset is quite balanced, which strengthens the confidence in the chosen weighted accuracy metric.

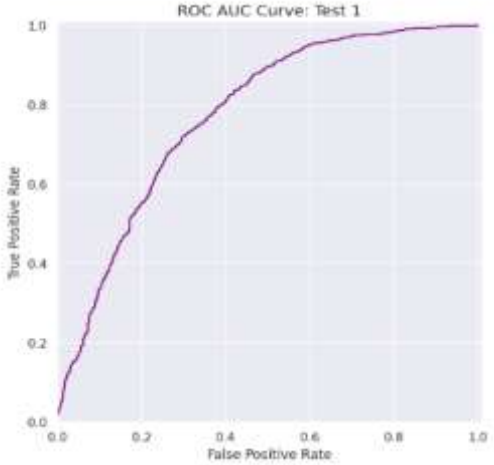


Fig 2.0 ROC Curve

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 7 | July 2024 - Peer Reviewed Journal

The second one is a Receiver Operating Characteristic (ROC) curve that shows the model's accuracy with regards to the classification under different threshold levels in terms of the True positive rate and False positive rate. This curve gives measures on how the model is able to classify samples between the positive and negative classes. The closer the curve is to the top left section the better the model performs, that's why the area intersects by 0,5. In this case, the ROC curve is quite favorable, and the position of the model when comparing it with other models, is close to the value of 1, as a result of the value of the area under the curve (AUC). 0. This in turns suggest that the model has high level of accuracy which allows for the differentiation of the two classes.

ROC is good for comparison of different models or for choosing threshold in the task where sensitivity and specificity is important. The nature of the curve shows how the true positive rates increase or decrease relating the false positive rate in order to evaluate the model's performance as it strives to achieve increased true positive and reduced false positives. An ideal model that has AUC at one point 1. Hence, any AUC of 0 is acceptable to represent the best or, in terms of classification, an AUC of 0 represents an A. 5 indicate no discriminative ability, which is equivalent to pure guess work

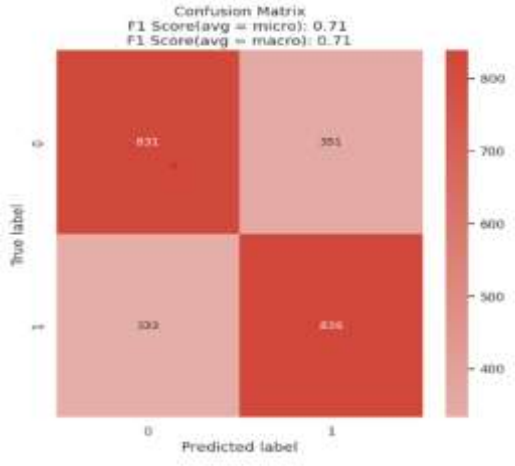


Fig 3.0 Confusion Matrix together with F1 Scores.

This is a matrix representing the confusion of a model with binary classification. This format has true labels in the vertical direction and the predicted labels in the horizontal axis with the values being the number of instances of true and predicted labels. By using the matrix, it can also be observed that the model got 831 instances wrongly classified as negative and 839 instances wrongly classified as positive. In the same regard, it incorrectly categorized 351 samples of negative class as belonging to a positive class, and 333 positive samples as belonging to a negative class. These misclassifications are essential when it comes to evaluating the model's accuracy, especially its precision and recall for each of the classes.

The F1 scores are also provided, and, similar to the case of the confusion matrix, these are the scores with the micro and macro average being 0. 71. The micro average F1 score calculates the total of true positives, false negatives, and false positives of all the classes to offer a single score on the performance. The macro average F1 score computes the F1 score of the classifier for every class individually to arrive at an average, where all classes are given importance at par. Both scores are equal that is at 0. It is 71 suggesting that the performance does not favor any of the classes.

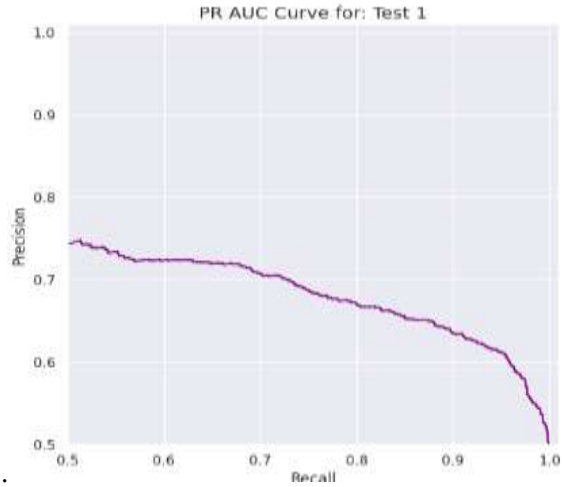


Fig 4.0 PR AUC Curve

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 7 | July 2024 - Peer Reviewed Journal

This one is an Area Under Precision-Recall (PR) curve that shows the Precision & Recall at various thresholds of the classification. Recall and precision are used as an x and y axis of a curve respectively. The curve is useful in indicating how the model's precision and recall metrics hold up when the threshold levels are adjusted. In this graph, the PR starts at nearly 1. 0 of precision and then falls as the recall increases. 8 and then falls as the value of recall rises implying that the model is most appropriate for precision at low recall and the reverse is true.

The area under the PR curve is a single scalar that gives an estimate of the performance of the classifier over all the threshold levels. Higher value of AUC represents better performance as the number of true positive is more with least number of true negative. The graph reveals that there's a negative correlation between the precision and the recall where the lower line depict the impact of the trade-off between the two as the precision declines to enhance the recall. This analysed that the designed model's precision is comparatively higher, when recall counts is comparatively lower, which means it can conveniently and properly identify the related instance; however, when it tries to yield better actual positive consequences (higher recall), it loses its capability of precision, suggesting it appropriately classifies more and more irrelevant instances.

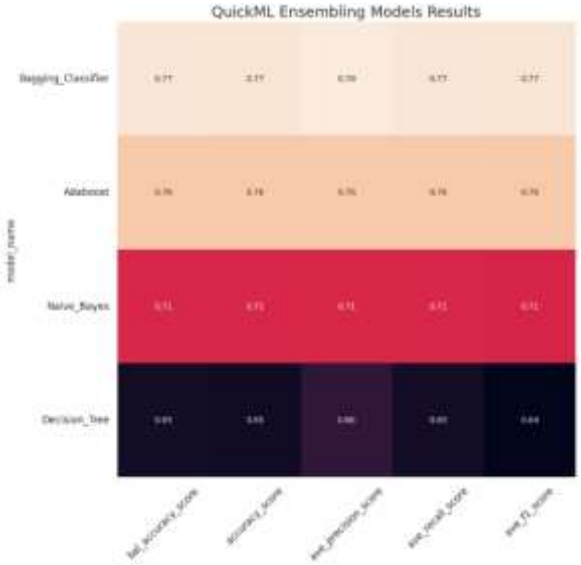


Fig 5.0 The Performance Heatmap

This graph is a heatmap that summarizes the performance of various machine learning models: From the list of classifiers, it introduced Bagging Classifier, AdaBoost Classifier, Naive Bayes, and Decision Tree. Here are the metrics that are depicted: balanced accuracy score, accuracy score, average precision recurring, average recall recurring and average F1 score. Thus, looking at the heatmap one can conclude that the Bagging Classifier has the highest overall accuracy with the scores of 0. 77 on most of the indices, which can be deemed as highly reliable and sustainable. AdaBoost is only slightly lower, gaining a score of 0. 76, Which shows that using combination of learners is efficient when implementing ensemble learning.

Nevertheless, Naive Bayes and Decision Tree classifiers are less efficient than that as they take relatively more time. The total scores of Naive Bayes for both classes are 0. 71, which Decision Tree has the worst performance depicted by presented scores of approximately 0. 65. Based on the findings presented above, it can be ascertained that methods that work in an ensemble such as Bagging and AdaBoost superior the single classifiers such as Naive Bayes and Decision Trees. The heatmap ensures that one is able to compare easily the models to determine which ones are most suitable for this given classification problem. Such a comparison is crucial when deciding on which model to use in production from a performance evaluation point of view.

8. CONCLUSION

In conclusion the work illustrates the approach to the problem of genetic disorders dataset which comprises data preprocessing, dealing with imbalanced data, model construction and model explanation. I also first had to encode the categorical features, Impute. Basic based on skewness and deleting outliers to have clean data. The above procedure is very important in ensuring that the models to be used perform as required and provide accurate results. In this respect, filtering away of outliers and the dealing of missing values are crucial in that they assist in data purification, and thus guarantee the models developed from the given data are both accurate and transportable.

To each testing target, the data was divided into the training and the testing datasets after which SMOTE technique was used to address the issue of imbalance, which is a crucial procedure when dealing with minority classes. AutoVIML was employed for constructing the models for predicting the property price automatically with feature selection and hyperparameters tuning for obtaining high accuracy. Also, a RandomForestClassifier was fit, and for interpretation, LIME is used which helps in getting features

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 7 | July 2024 - Peer Reviewed Journal

that the model focuses on while making a decision. This is significant in the aspect of model trust due to the nature of the scores provided, which pertain to probabilities of genetic disorders, among others. This makes sure that the analysis is realized with the correct value and it is also understandable and useful.

9. REFERENCES

1. W. MCKINNEY, AND OTHERS "PANDAS: THIS IS A CORE PYTHON LIBRARY FOR DATA ANALYSIS AND STATISTICS. JSS: JOURNAL OF OPEN SOURCE SOFTWARE, VOLUME 3, NO. 29, 2018.
2. PEDREGOSA, F. , VAROQUAUX, G. , GRAMFORT, A. , MICHEL, V. , THIRION, B. , GRISEL, O. , . . . & DUCHESNAY, É. "SCIKIT-LEARN: UNDERSTANDING OF "MACHINE LEARNING IN PYTHON". ACM JOURNAL OF MACHINE LEARNING RESEARCH; VOLUME. 12, PP. 2825-2830, 2011.
3. CHAWLA, N. V. , BOWYER, K. W. , HALL, L. O. , & KEGELMEYER, W. P. "SMOTE: OTHERWISE REFERRED TO AS SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE. JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH VOLUME NUMBER. 16, PP. 321-357, 2002.
4. BREIMAN L. "RANDOM FORESTS". MACHINE LEARNING, VOL. 45, NO. 1, PP. 5–32, 2001.
5. SHANKAR, S. , LORIA, S. , "AUTO_VIML: IT LAMENTS HOW THE FIELD COMMONLY REFERS TO COMPLEX MACHINE LEARNING AS "AUTOMATIC VARIENT INTERPRETABLE MACHINE LEARNING".
6. RIBEIRO, M. T. , SINGH, S. , & GUESTRIN, C. ; "WHY SHOULD I TRUST YOU? EXPLAINING THE PREDICTIONS OF ANY CLASSIFIER". IN KDD 2016, THE 22ND ACM SIGKDD CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, ACM, NEW YORK, NY, USA, PP. 1135-1144.
7. HAN, H. , WANG, W. , & MAO, B. "BORDERLINE-SMOTE: "A NEW OVER-SAMPLING METHOD IN IMBALANCED DATA SETS LEARNING". PUBLISHED IN THE BOOK ''PROCEEDINGS OF THE 2005 INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTING'' ON PP 878-887, 2003.
8. LIAW, A. , & WIENER, M. 'CLASSIFICATION AND REGRESSION BY RANDOMFOREST'. R NEWS, VOL. PDF FROM WWW. NARIC-NIDDK. ORG RESEARCH DIVISION, 2, NO. 3, PP. 18-22, 2002.
9. VAN DER WALT, S. , COLBERT, S. C. , & VAROQUAUX, G. "THE NUMPY ARRAY: A PROVIDING A STRUCTURE FOR EFFICIENT NUMERICAL COMPUTATION. COMPUTING IN SCIENCE & ENGINEERING, VOLUME. 13, NO. 2, PP. 22-30 DOI:10. 4148/0148-6077. 1344 2011.
10. HUNTER, J. D. "MATPLOTLIB: SO, HERE IS PRESENTED A SUBSYSTEM OF A PROGRAM, TERMED "A 2D GRAPHICS ENVIRONMENT". IEEE-CS USENIX ASSOCIATION COMPUTING IN SCIENCE AND ENGINEERING VOLUME. 9, NO. 3, PP. 90-95, OCTOBER 2007.